Advanced Regression: 2c Distributed non-linear lag models and other extensions

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

28th February 2023

Overview

Concepts we cover in this lecture:

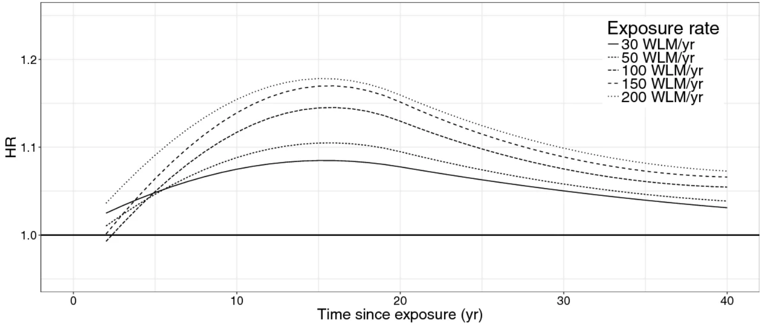
- Distributed lag non-linear models
- Cross-basis function
- Case studies

Introduction of the problem

- An exposure event is frequently associated with a risk lasting for a defined period in the future
- The risk at a given time is assumed a result of protracted exposures experienced in the past
- Examples include, drugs, carcinogens, etc.

Challenge: The risk should be modelled in terms of contributions depending on intensity and timing of the exposure events: bi-dimensional association (interaction)

Example 1: Lung cancer and radon exposure

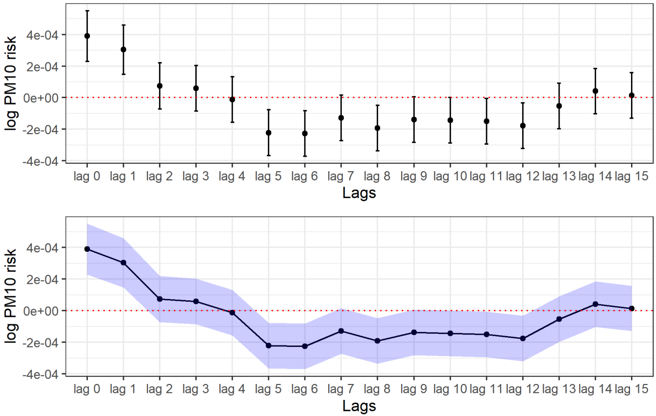


Example 2: PM10 in Chicago

```
k <- 1:16
res_store <- list()

for(i in 1:length(k)){
chicagoNMMAPS$pm10_laggeg <- lag(chicagoNMMAPS$pm10, n = k[i]-1)

mgcv::gam(death ~ s(temp) +
        s(time) + s(month) + dow + pm10_laggeg,
        data = chicagoNMMAPS, family = "poisson") -> tmp

res_store[[i]] <- list(est = coef(tmp)["pm10_laggeg"] ,
        LL = coef(tmp)["pm10_laggeg"] - 1.96*summary(tmp)$se["pm10_laggeg"],
        UL = coef(tmp)["pm10_laggeg"] + 1.96*summary(tmp)$se["pm10_laggeg"])
}

lapply(res_store, unlist) %>% do.call(rbind, .) %>% as_tibble() %>% mutate(type =
        factor(paste0("lag_", 0:15), levels = paste0("lag_", 0:15))) -> plotres

ggplot(data = plotres) +
        geom_point(aes(x=type, y=est.pm10_laggeg)) +
        geom_errorbar(aes(x=type, ymin=LL.pm10_laggeg, ymax=UL.pm10_laggeg, width = 0.1)
        geom_hline(yintercept = 0, col = "red", linetype = "dotted") + theme_bw() +
        ylab("log_risk_PM10") + xlab("Lags") |

ggplot(data = plotres) +
        geom_point(aes(x=type, y=est.pm10_laggeg)) +
        geom_line(aes(x=type, y=est.pm10_laggeg, group=1)) +
        geom_ribbon(aes(x=type, ymin=LL.pm10_laggeg, ymax=UL.pm10_laggeg, group = 1),
        fill = "blue", alpha = 0.2) + geom_hline(yintercept = 0, col = "red",
        linetype = "dotted") + theme_bw() + ylab("log_risk_PM10") + xlab("Lags")
```

Example 2: PM10 in Chicago

- Which are the main assumptions here?



Distributed linear lag models

The unconstrained distributed lag model of order q is:

$$Y_t = \beta_0 + \beta_{10} X_t + \beta_{11} X_{t-1} + \cdots + \beta_{1q} X_{t-q} + \epsilon_t$$

- $\beta_{1\ell}$ is the effect at lag $\ell = 0, 1, \ldots q$ and ϵ_t an error term.
- The overall impact for a unit change in X is given by $\sum_{\ell=0}^{q} \beta_\ell$.

Example 2: PM10 in Chicago

```
chicagoNMMAPS$pm10_laggeg0 <- lag(chicagoNMMAPS$pm10, n = 0)
chicagoNMMAPS$pm10_laggeg1 <- lag(chicagoNMMAPS$pm10, n = 1)
chicagoNMMAPS$pm10_laggeg2 <- lag(chicagoNMMAPS$pm10, n = 2)
chicagoNMMAPS$pm10_laggeg3 <- lag(chicagoNMMAPS$pm10, n = 3)

mgcv::gam(death ~ s(temp) +
        s(time) + s(month) + dow + pm10_laggeg0 + pm10_laggeg1 + pm10_laggeg2 +
        pm10_laggeg3, data = chicagoNMMAPS, family = "poisson") %>% summary()

Formula:
death ~ s(temp) + s(time) + s(month) + dow + pm10_laggeg0 + pm10_laggeg1 +
pm10_laggeg2 + pm10_laggeg3

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.707e+00  5.825e-03 808.083  < 2e-16 ***
dowMonday      2.898e-02  5.375e-03   5.391 7.01e-08 ***
dowTuesday     2.326e-02  5.428e-03   4.285 1.82e-05 ***
dowWednesday   5.054e-03  5.447e-03   0.928 0.353471
dowThursday    5.898e-03  5.385e-03   1.095 0.273448
dowFriday      1.294e-02  5.336e-03   2.426 0.015264 *
dowSaturday    1.931e-02  5.273e-03   3.661 0.000251 ***
pm10_laggeg0   3.958e-04  8.997e-05   4.399 1.09e-05 ***
pm10_laggeg1   1.765e-04  9.259e-05   1.907 0.056571 .
pm10_laggeg2  -2.595e-05  9.039e-05  -0.287 0.774093
pm10_laggeg3   1.188e-04  8.350e-05   1.423 0.154769
```

Considerations

- Easy implementation when lags are few; overparametrized when we want to assess a lot of lags
- Collinearity issues: The exposure is likely to be highly correlated with the values of the previous/after days. Weird behaviours in the point estimates (surprising protective effects), variance inflation.

Alternative: to impose some constraints:

- A constant effect within lag intervals
- Average of the exposures in the previous L day
- Describing the coefficients with a smooth curve using continuous functions such as splines, polynomials, and other basis functions.

The idea: β_ℓ can be modelled using a basis function.

Polynomial DLM

Let $\beta_\ell = \sum_j^p \tau_j \ell^j, \quad \ell = 0, \ldots, q$, lets write it for 2 lags using a 3rd degree polynomial to see it explicitly:

$$Y_t = \beta_0 + \beta_{10} X_t + \beta_{11} X_{t-1} + \beta_{12} X_{t-2} + \epsilon_t$$

$$\beta_{10} = \tau_0,\ \beta_{11} = \tau_0 + \tau_1 + \tau_2 + \tau_3,\ \beta_{12} = \tau_0 + \tau_1 2 + \tau_2 2^2 + \tau_3 2^3$$

and we can modify as per first lecture to model more localized structures using: $\beta_\ell = \sum_j^p \tau_j \ell^j + \sum_k^K \nu_k (\ell - \kappa_k)_+^p$, thus:

$$\beta_{10} = \tau_0 + \nu_1 (0 - \kappa_1)_+^3 + \cdots + \nu_K (0 - \kappa_K)_+^3,$$
$$\beta_{11} = \tau_0 + \tau_1 + \tau_2 + \tau_3 + \nu_1 (1 - \kappa_1)_+^3 + \cdots + \nu_K (1 - \kappa_K)_+^3,$$
$$\beta_{12} = \tau_0 + \tau_1 2 + \tau_2 2^2 + \tau_3 2^3 + \nu_1 (2 - \kappa_1)_+^3 + \cdots + \nu_K (2 - \kappa_K)_+^3$$

and similarly we can penalize it can estimate the *penalised spline distributed lag estimate* of β_ℓ

Polynomial DLM in R: Chicago

```
cb1.pm <- crossbasis(chicagoNMMAPS$pm10, lag=15, argvar=list(fun="lin"),
        arglag=list(fun="poly", degree=4))

summary(cb1.pm)

CROSSBASIS FUNCTIONS
observations: 5114
range: -3.049835 to 356.1768
lag period: 0 15
total df:  5

BASIS FOR VAR:
fun: lin
intercept: FALSE

BASIS FOR LAG:
fun: poly
degree: 4
scale: 15
intercept: TRUE

model_dlm <- mgcv::gam(death ~ s(temp) + s(time) + s(month) + dow + cb1.pm,
        family=poisson(), chicagoNMMAPS)

summary(model_dlm)

pred1.pm <- crosspred(cb1.pm, model_dlm, at=0:20, bylag=0.2)
plot(pred1.pm, ptype = "slices", var = 1, cumul=FALSE, ylab="RR",
         main="Association with a 1-unit increase in PM10")
```

Polynomial DLM in R: Chicago

```
Family: poisson
Link function: log

Formula:
death ~ s(temp) + s(time) + s(month) + dow + cb1.pm

Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.734e+00  1.028e-02 460.746  < 2e-16 ***
dowMonday     3.162e-02  5.933e-03   5.330 9.85e-08 ***
dowTuesday    2.100e-02  5.998e-03   3.501 0.000463 ***
dowWednesday  3.579e-03  6.050e-03   0.592 0.554073
dowThursday   3.367e-03  6.069e-03   0.555 0.579092
dowFriday     1.339e-02  6.054e-03   2.212 0.026984 *
dowSaturday   1.805e-02  5.957e-03   3.031 0.002439 **
cb1.pmv1.l1   3.062e-04  7.862e-05   3.895 9.81e-05 ***
cb1.pmv1.l2  -2.115e-03  1.068e-03  -1.979 0.047789 *
cb1.pmv1.l3   3.966e-03  4.423e-03   0.897 0.369884
cb1.pmv1.l4  -3.477e-03  6.698e-03  -0.519 0.603653
cb1.pmv1.l5   1.348e-03  3.323e-03   0.406 0.684882

Approximate significance of smooth terms:
edf Ref.df Chi.sq p-value
s(temp)  8.584  8.940  165.0  <2e-16 ***
s(time)  7.658  8.530  261.1  <2e-16 ***
s(month) 8.125  8.806  278.3  <2e-16 ***

R-sq.(adj) =  0.267   Deviance explained = 29.1%
```

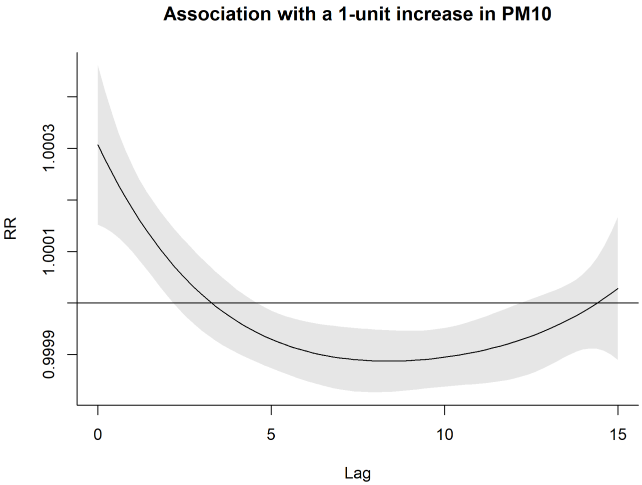
Polynomial DLM in R: Chicago

- Retrieve the cumulative effect. What is the interpretation here?

```
> pred1.pm$allRRfit["1"]
1
0.9997201
> pred1.pm$allRRlow["1"]
1
0.9991616
> pred1.pm$allRRhigh["1"]
1
1.000279
```

Polynomial DLM in R: Chicago

What is the main assumption here? Can we relax it?

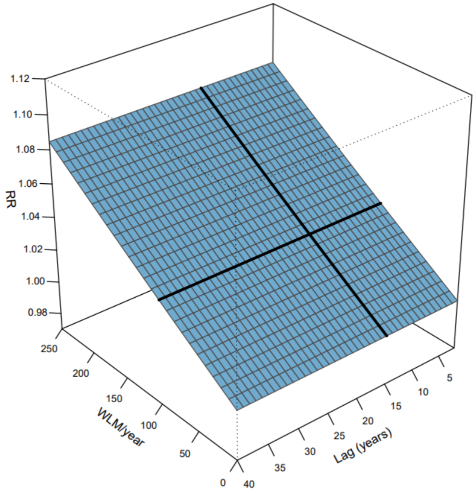


Extension to distributed non-linear lag models

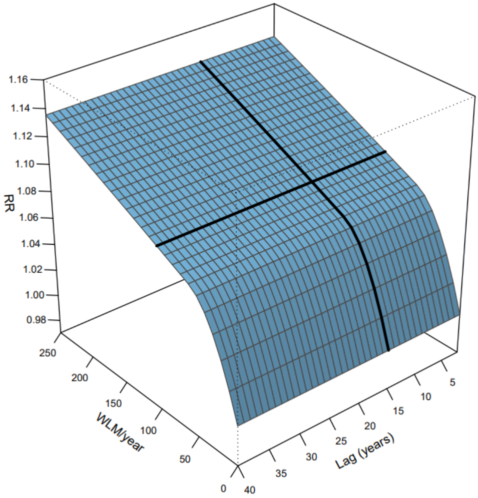
- We know that temperature and mortality have a U-shape relationship
- We know that high temperature has a lag effect on mortality
- Can we define models to combine these two components?

The idea: to calculate this bi-dimensional relationship, we need a basis function that combines the basis function in the lag dimension and the basis function in the exposure dimension: **Cross-basis function**

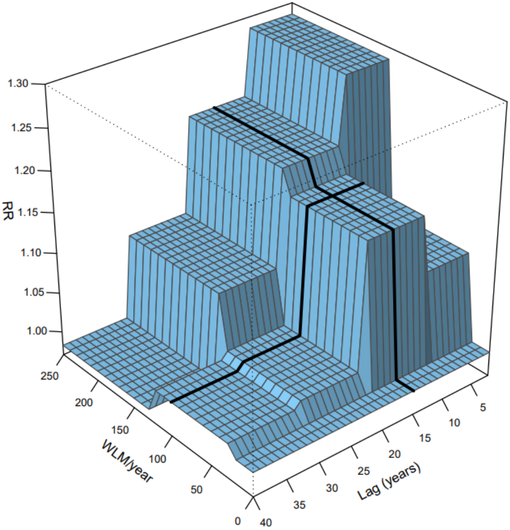
linear-by-constant



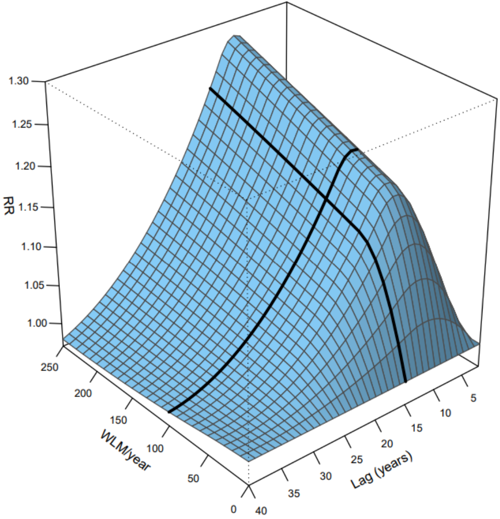
spline-by-constant



step-by-step


RR
1.30
1.25
1.20
1.15
1.10
1.05
1.00
WLM/year
0
50
100
150
200
250
Lag (years)
5
10
15
20
25
30
35
40

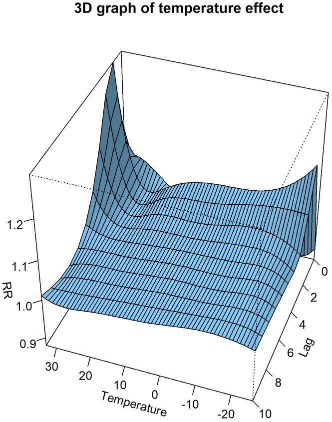
spline-by-spline

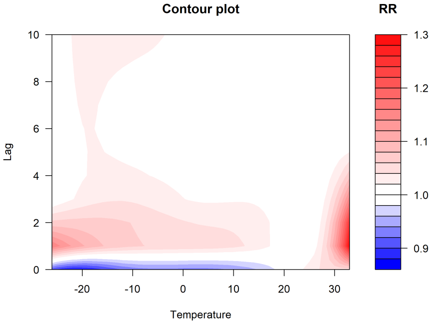


Example 3: Temperature in Chicago

```
cb2.pm <- crossbasis(chicagoNMMAPS$pm10, lag=1, argvar=list(fun="lin"),
        arglag=list(fun="strata"))

varknots <- equalknots(chicagoNMMAPS$temp, fun="bs", df=5, degree=2)
lagknots <- logknots(10, 3)
cb2.temp <- crossbasis(chicagoNMMAPS$temp, lag=10, argvar=list(fun="bs",
        knots=varknots), arglag=list(knots=lagknots))


model_dlm2 <- mgcv::gam(death ~ cb2.pm + cb2.temp + s(time) + s(month) + dow,
        family=poisson(), chicagoNMMAPS)

pred2.temp <- crosspred(cb2.temp, model_dlm2, cen=21, by=1)

plot(pred2.temp, xlab="Temperature", zlab="RR", theta=200, phi=40, lphi=100,
        main="3D graph of temperature effect")

plot(pred2.temp, "contour", xlab="Temperature", key.title=title("RR"),
plot.title=title("Contour plot",xlab="Temperature",ylab="Lag"))
```

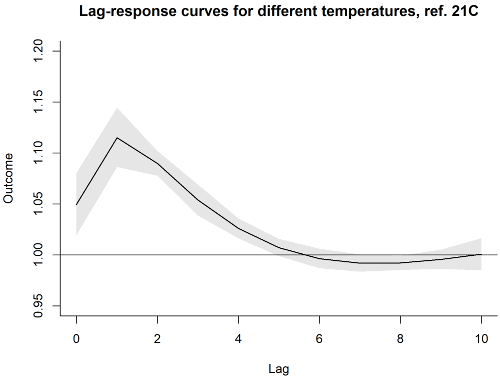
Example 3: Temperature in Chicago





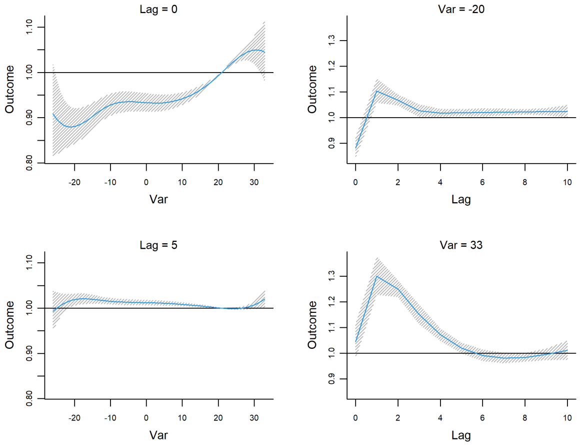
Example 3: Temperature in Chicago

```
plot(pred2.temp, "slices", var=30, col=1, ylim=c(0.95,1.2), lwd=1.5,
main="Lag-response curves for different temperatures, ref. 21C")
```



Example 3: Temperature in Chicago

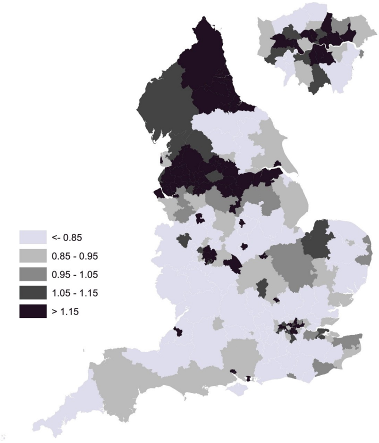
```
plot(pred2.temp, "slices", var=c(-20,33), lag=c(0,5), col=4,
ci.arg=list(density=40,col=grey(0.7)))
```



Example 4: Extension to space

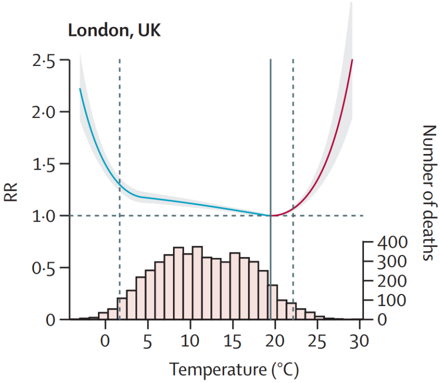
Warm temperatures and COPD hospitalisations in England: A nationwide case-crossover study during 2007-2018.

- 3$^{\text{rd}}$ cause of death, 3.17 million deaths in 2015 globally.
- In England, 115,000 emergency admissions and 24,000 deaths per year.
- COPD exacerbations: Bacteria, viruses and air-pollution.
- The role of temperature is unclear.



Temperature

- Typically U-shaped relationship between temperature and health.
- Cold, dry air or hot air can trigger a flare-up.
- Different confounding, different lags across different temperatures.
- This study focuses on warm temperatures.



Previous studies

Authors	Aggregation	Country	Pollutants	Effect
Michelozzi 2009 et al	city & daily	EU	NO_2, O_3	2.1 (0.6 to 3.6) per 1°C
Anderson et al 2013	county & daily	US	O_3, PM_{10}, $PM_{2.5}$	2.0 (0.4, 4.5) per 10°F
Zhao 2019 et al	individual	Brazil	no adjustment	5.0 (4.0, 6.0) per 5°C

- Spatial & temporal aggregation
  - Exposure varies on high resolution.
  - Insufficient adjustment for confounding (for instance physical activity).
  - Ecological bias
- One study individual data, but did not adjust for air-pollution

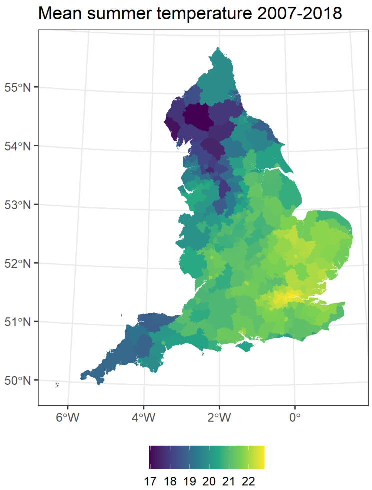
Outcome and Exposure

Outcome

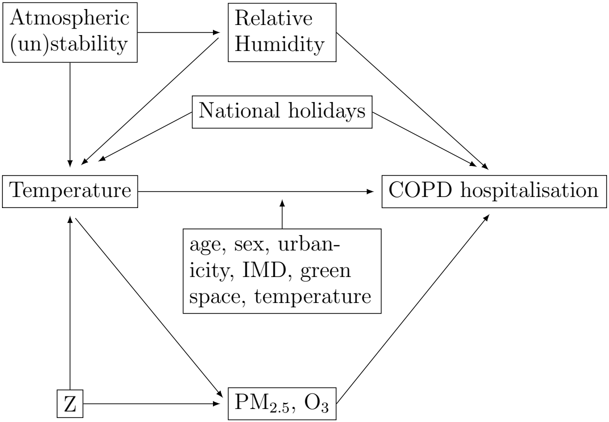
- NHS digital & SAHSU.
- COPD hospitalization (ICD10 J40-44) 2007-2018.
- Individual data/ 100m grid spatial resolution.
- Summer months.

Exposure

- Daily maximum temperature 2007-2018 at 1km grid from MetOffice.
- lag0-2.



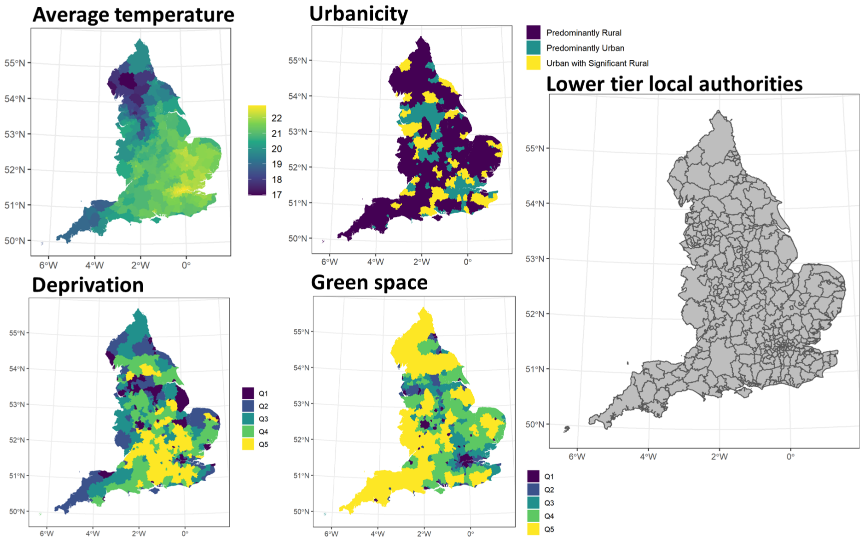
Confounding



Covariates

Covariates	Source	Space	Time	years
$PM_{2.5}$	MetOffice	$1km^2$	daily	2007-2018
O_3	MetOffice	$1km^2$	daily	2007-2018
Relative humidity	MetOffice	$10km^2$	daily	2007-2018
Holidays	ONS	nationwide	daily	2007-2018

Spatial effect modifiers



Step 1. Find linear threshold

Let Y_{ij} be an indicator of the COPD hospitalization at time i (1-case, 0-control) of the j-th group of cases-controls, and μ_{ij} the risk ratio:

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ijk}) &= \alpha_1 I(X_{1i} < c_l)X_{1i} + \alpha_2 I(X_{1i} \geq c_l)X_{1i} + \\
&\qquad \sum_{m=1}^{4} \beta_m Z_{mi} + u_j + w_k \\
u_j &\sim N(0, 100) \\
w_k &\sim N(0, \sigma^2) \\
\alpha_1, \alpha_2, \beta_1, \ldots \beta_4 &\sim N(0, 1) \\
\sigma &\sim \text{Gamma}(p, q)
\end{aligned}
$$

Step 2a. Effect modification by age and sex

We fitted the previous model for c_* that minimizes the WAIC for the different sex and age group (<65, 65−85, >85) g subgroups and patient k.

$$
\begin{aligned}
Y_{ijgk} &\sim \text{Poisson}(\mu_{ijgk}) \\
\log(\mu_{ijgk}) &= \alpha_1 I(X_{1ig} < c_*)X_{1ig} + \alpha_2 I(X_{1ig} \geq c_*)X_{1igk} + \\
&\qquad \sum_{m=1}^{4} \beta_m Z_{mig} + u_j + w_k \\
u_j &\sim N(0, 100) \\
w_k &\sim N(0, \sigma^2) \\
\alpha_1, \alpha_2, \beta_1, \ldots \beta_5 &\sim N(0, 1)
\end{aligned}
$$

Step 2b. Spatial Effect modification

$$Y_{ijk} \sim \text{Poisson}(\mu_{ijk})$$

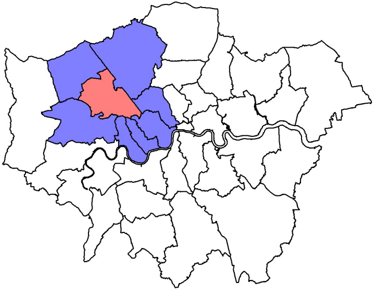
$$\log(\mu_{ijk}) = \alpha_1 I(X_{1i} < c_*)X_{1i} + \alpha_{2s} I(X_{1i} \geq c_*)X_{1i} + \sum_{m=1}^{4} \beta_m Z_{mi} + u_j + w_k$$

$$\alpha_{2s} = \alpha_2 + \sum_{m=1}^{8} \gamma_m H_{sm} + v_s + b_s$$

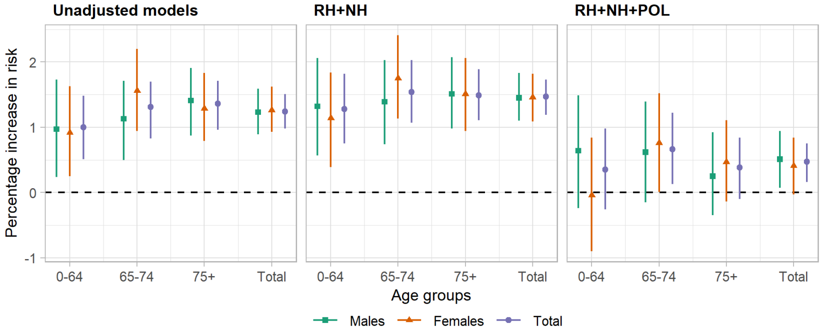
$$w_k \sim N(0, \sigma_1^2)$$

$$v_s \sim N(0, \sigma_2^2)$$

$$b_s | b_{-s} \sim N\left(\frac{\sum_{s \sim r} w_{rs} b_s}{\sum_{s \sim r} w_{rs}}, \frac{\sigma_2^3}{\sum_{s \sim r} w_{rs}} \right)$$

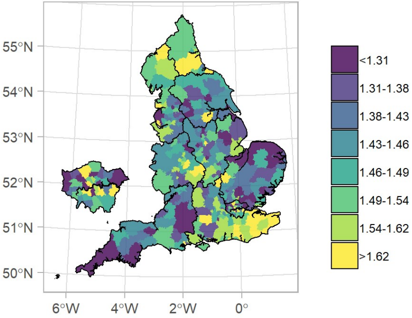


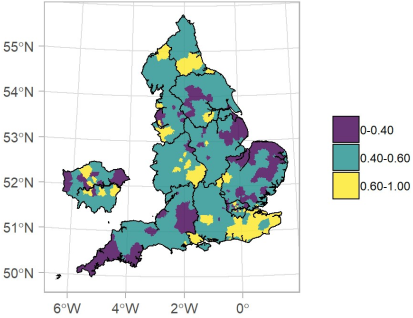
Step 2a: Effect modification by age and sex



Step 2a: Spatial effect modification

Results unadjusted for spatial effect modifiers



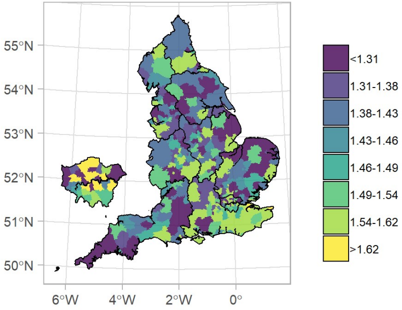


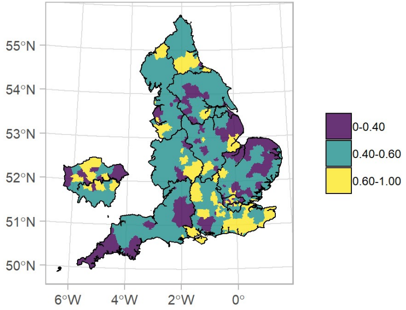
Step 2b: Spatial effect modification

Effect modifier	Percentage increase	Pr(Covariate>0)
Green space	-1.46 (-6.99, 4.39)	0.30
Average temperature	-0.41 (-1.49, 0.71)	0.22
IMD		
Q1	1	
Q2	0.81 (-1.16, 3.08)	0.78
Q3	1.57 (-0.76, 4.06)	0.91
Q4	0.75 (-1.68, 3.36)	0.71
Q5	1.62 (-1.31, 4.49)	0.85
Predominantly Rural	1	
Urban with significant rural	-0.79 (-3.10, 1.51)	0.25
Predominantly urban	-1.57 (-4.16, 0.96)	0.12

Step 2a: Spatial effect modification

Results adjusted for spatial effect modifiers





Summary of the results

- Unadjusted: 1.2% (-1.0%, 1.5%) for every 1^{o}C increase in warm temperatures.
- Adjusted: 1.5% (1.2%, 1.7%) for every 1^{o}C increase in warm temperatures.
- Weak evidence of an effect modification by sex and age.
- Strong spatial effect modification, with some evidence that populations in areas with more green space, higher average temperature and urbanicity are least vulnerable.

Conclusion

- Evidence COPD hospital admissions and maximum temperatures higher than 23.8°C during the summer months.
- Spatial vulnerabilities partly can be explained by green space, deprivation, urbanicity and average temperature.

Take home message

Evidence that COPD hospitalisations increase with warmer temperatures and as temperatures consistently increase, public health systems should be alerted and prepared to challenge the increased COPD hospitalisation burden.

`https://github.com/gkonstantinoudis/COPDTempSVC`

Summary

- Extent basis function to incorporate the different lags
- Distributed lag linear models
- Distributed lag non-linear models
- Extension in the spatial dimension.

Check: `https://cran.r-project.org/web/packages/dlnm/vignettes/dlnmTS.pdf`
Questions?